Multi-class detection of cherry tomatoes using improved YOLOv4-Tiny

Fu Zhang[1,2*], Zijun Chen[1], Shaukat Ali[3], Ning Yang[4], Sanling Fu[5], Yakun Zhang[1]

(1. *College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang 471003, Henan, China*;
2. *Collaborative Innovation Center of Machinery Equipment Advanced Manufacturing of Henan Province, Luoyang 471003, Henan, China*;
3. *Wah Engineering College, University of Wah, Wah Cantt 47040, Pakistan*;
4. *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China*;
5. *College of Physics Engineering, Henan University of Science and Technology, Luoyang 471023, Henan, China*)

Abstract: The rapid and accurate detection of cherry tomatoes is of great significance to realizing automatic picking by robots. However, so far, cherry tomatoes are detected as only one class for picking. Fruits occluded by branches or leaves are detected as pickable objects, which may cause damage to the plant or robot end-effector during picking. This study proposed the Feature Enhancement Network Block (FENB) based on YOLOv4-Tiny to solve the above problem. Firstly, according to the distribution characteristics and picking strategies of cherry tomatoes, cherry tomatoes were divided into four classes in the nighttime, and daytime included not occluded, occluded by branches, occluded by fruits, and occluded by leaves. Secondly, the CSPNet structure with the hybrid attention mechanism was used to design the FENB, which pays more attention to the effective features of different classes of cherry tomatoes while retaining the original features. Finally, the Feature Enhancement Network (FEN) was constructed based on the FENB to enhance the feature extraction ability and improve the detection accuracy of YOLOv4-Tiny. The experimental results show that under the confidence of 0.5, average precision (AP) of non-occluded, branch-occluded, fruit-occluded, and leaf-occluded fruit over the day test images were 95.86%, 92.59%, 89.66%, and 84.99%, respectively, which were 98.43%, 95.62%, 95.50%, and 89.33% on the night test images, respectively. The mean Average Precision (mAP) of four classes over the night test set was higher (94.72%) than that of the day (90.78%), which were both better than YOLOv4 and YOLOv4-Tiny. It cost 32.22 ms to process a 416×416 image on the GPU. The model size was 39.34 MB. Therefore, the proposed model can provide a practical and feasible method for the multi-class detection of cherry tomatoes.

Keywords: cherry tomatoes, deep learning, data augmentation, YOLOv4, occlusion, multi-class detection
DOI: 10.25165/j.ijabe.20231602.7744

Citation: Zhang F, Chen Z J, Ali S, Yang N, Fu S L, Zhang Y K. Multi-class detection of cherry tomatoes using improved YOLOv4-Tiny. Int J Agric & Biol Eng, 2023; 16(2): 225–231.

1 Introduction

As the subsystems of agricultural picking robots, the vision system can identify and locate fruit accurately in a natural growth environment, and it can guide the robot picking end-effector to smoothly grasp and separate the fruit from the plant to complete picking tasks[1-3]. The cherry tomato is a variety of tomatoes that included higher antioxidants and phytochemical compounds in the greenhouse[4]. With the growth of domestic and foreign market demand, how to detect quickly and accurately cherry tomatoes in greenhouse environment is particularly important.

Traditional image processing combined with machine learning methods has been widely studied in fruit and vegetable recognition[5-8], but the detection accuracy is limited by light, complex backgrounds, and environmental information[9-11]. With the development of smart agriculture, deep learning has shown significant advantages in fruit detection[12-15]. Zhang et al.[16] proposed a lightweight YOLOv4 for cherry tomato detection. Xu et al.[17] employed the YOLOv3-Tiny to detect tomatoes and achieved an F1 score of 91.92%. Zhang et al.[18] used Faster R-CNN to recognize tomatoes in complex environments and achieved the average correct rate of 95.2%. Considering the clustering growth habit of cherry tomatoes and improving the picking efficiency, Xu et al.[19] and Zhang et al.[20] developed the identification algorithm and the picking robotic manipulator related to the tomato bunch harvest.

Deep learning has achieved accurate and rapid detection of tomatoes. Although scholars considered the influence of occlusion, cherry tomatoes are still classified into one class that could be picked. The end-effector grabs them together if fruits occluded by branches or leaves are identified as direct picking objects. It poses a significant risk of end-effector and plant damage during harvesting. Therefore, fruits under different occluded conditions should be distinguished at the time of identification. Gao et al.[21] classified apples in the SNAP system into four classes including non-occluded, leaf-occluded, branch/wire-occluded, and fruit-occluded fruits, and these four classes of apples were detected based on Faster Regional Convolutional Neural Network (Faster RCNN). According to the picking strategy, Suo et al.[22] divided kiwi fruit into five classes and YOLOv4 had the highest mAP of 91.9% by comparing with YOLOv3.

Received date: 2022-06-17 **Accepted date:** 2023-02-21

Biographies: Zijun Chen, Master candidate, research interest: agricultural informatization and detection technology, Email: rssnczj98@163.com; **Shaukat Ali**, Assistant Professor, Chairperson, Wah Engineering College, University of Wah, research interest: intelligent equipment and technology, Email: shaukat.ali@wecuw.edu.pk; **Ning Yang**, PhD, Professor, research interest: microsensor and automatic detection technology, Email: yangn@ujs.edu.cn; **Sanling Fu**, PhD, Assistant Professor, research interest: intelligent measurement technology, Email: fusanling@126.com; **Yakun Zhang**, PhD, Lecturer, research interest: intelligent measurement technology, Email: zhangyk@haust.edu.cn.

***Corresponding author: Fu Zhang**, PhD, Professor, Assistant Dean, research interest: agricultural informatization and intelligent agricultural equipment technology. College of Equipment Agricultural Engineering, Henan University of Science and Technology, Luoyang, 471003, China. Tel: +86-379-65627650, Email: zhangfu30@126.com.

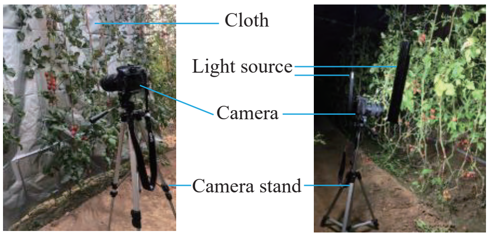
This research believed that cherry tomatoes can also be divided into multiple classes for detection based on the above research results. Therefore, a deep learning model based on a widely used model YOLOv4-Tiny was proposed to implement multi-class detection of cherry tomatoes. The CSPNet with the hybrid attention mechanism was used to build the Feature Enhancement Network Block (FENB) to achieve effective feature extraction. The Feature Enhancement Network (FEN) was constructed based on the feature enhancement module to improve the performance of YOLOv4-Tiny.

2 Materials and methods

2.1 Data set

2.1.1 Image data collection

The image acquisition location of cherry tomatoes in this study was in a greenhouse in Mengjin district, Luoyang city, Henan province, China. The acquisition times were 9:00-11:00 a.m., 2:00-4:00 p.m., and 7:00-11:00 p.m. on November 27, 2021. The Canon camera (Canon EOS 750D, Japan), a camera mount, two artificial strip Light Emitting Diodes (LEDs), and a light source digital-analog controller were used to collect cherry tomatoes images, as shown in Figure 1.



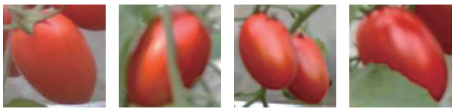
a. Day and double row b. Night and single row

Figure 1 Cherry tomatoes images collection

The camera was 30-50 cm from the tomato plants, which was beneficial to reduce background complexity to manage the growth environment under artificial intervention. The experimenter placed a white cloth between the cherry tomato plants and the adjacent row of plants (Figure 1a) to improve the visibility of the fruit. No cloth was placed for single-row plants (Figure 1b). The two LEDs created lighting conditions at the night. A total of 1103 RGB images (596 of the day and 507 of the night) were obtained with a resolution of 4000 pixels×4000 pixels. All images were saved in '.jpg' format.

2.1.2 Image data set

Classification of cherry tomatoes can be divided into four classes under different occlusion conditions as shown in Figure 2. The first class of cherry tomatoes is not occluded (NO), which can be picked directly by the robot end-effector. The second class is occluded by branch (OB) which is defined as being unable to be picked. The third class is occluded by other fruits (OF), which can be picked in order from the outside one to the inside. The fourth class is occluded by leaves (OL), where the robot end-effector pokes leaves to grab. Therefore, the vision system of the picking robot can identify and detect the above-mentioned four classes of cherry tomatoes in the greenhouse, which can allow the robot to choose correspondingly picking strategies.



a. NO b. OB c. OF d. OL

Note: NO: Cherry tomato is not occluded; OB: Cherry tomato is occluded by branch; OF: Cherry tomato is occluded by other fruits; OL: Cherry tomato is occluded by leaves. Same below.

Figure 2 Classification of cherry tomatoes under different occlusion conditions

The image data set flowchart of Multi-class cherry tomatoes was shown in Figure 3. The obtained 4000×4000 images were scaled to 416×416 images in order to train the model more quickly. Then, the LabelImg software was used to manually annotate cherry tomatoes into the above four classes. To enhance the generalization ability of the model and to avoid overfitting, the images were preprocessed in terms of rotation, mirroring, color balance, blurring, brightness, and adding noise to enrich the samples. After data augmentation, the images in daytime were enlarged from 596 to 9536, and the images in nighttime were enlarged from 507 to 8112. It is worth noting that the '.xml' files corresponding to each image would also be augmented accordingly. The daytime and night data sets were established and randomly divided into the training set, validation set, and test set according to the ratio of 6:2:2, respectively. The operations run under Windows 10 and python 3.7.

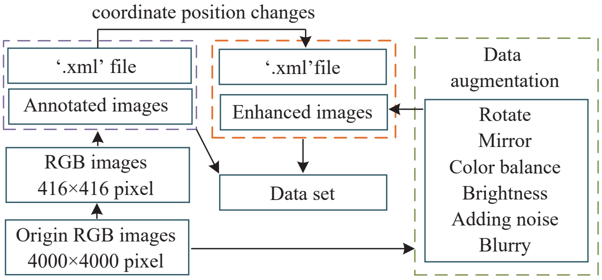


Figure 3 Image data set flowchart of multi-class cherry tomatoes

2.2 Algorithm

Compared with two-stage detection models[23], the YOLO (You Only Look Once) series show the characteristics of high efficiency and flexibility performance[24]. Among them, YOLOv4-Tiny[25] is a lightweight object detection model, which is more suitable for the deployment of mobile terminals or embedded devices to meet the requirements of the agricultural robot visual systems.

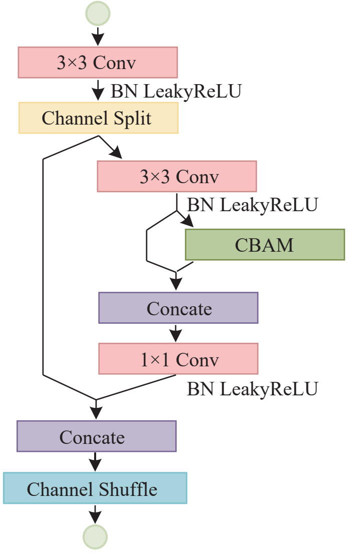
2.2.1 YOLOv4-Tiny

YOLOv4-Tiny consists of three parts, backbone network, neck network, and head network[26]. The backbone network CSPDarknet53_tiny consists of two stacked CBL modules, three Resblocks, and a CBL module. Leaky-ReLU was used as the activation function to improve computational speed[27]. The neck network is the feature fusion network using Feature Pyramid Networks (FPN), where the model strengthens the target features with different scales to further detect the target. The head network[28] utilizes the two feature layers obtained by FPN for object prediction.

2.2.2 Feature Enhancement Network Block

CSPDarknet53_tiny of YOLOv4-Tiny gets faster speed with fewer parameters reducing model detection accuracy. This study proposed CSPNet[29] with the hybrid attention mechanism to construct Feature Enhancement Network Block (FENB) for improving recognition accuracy. The structure of the FENB is shown in Figure 4.

FENB employs the CBL with the first convolution kernel size of 3×3 to extract global features, then divides the channel into two parts evenly, where the second part is taken to reduce memory traffic. The second part flows into Path A and Path B, respectively to increase the gradient path. Next, the second convolution kernel

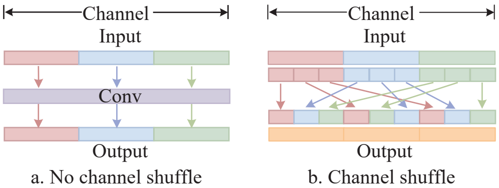


Note: Conv: Convlution; BN: Batch Normalization; LeakyReLU: Leaky Rectified Linear Unit; CBAM: Convolutional Block Attention Module.

Figure 4 Diagram schematic of Feature Enhancement Network Block structure

size of 3×3 CBL is used to integrate features further to enrich feature-level information. Then, the Convolutional Block Attention Module[30] (CBAM) extracts the effective information in the feature map of the inflow path C, which makes FENB pay more attention to identifying the target object. Finally, the feature map integrated by the convolution kernel size of 1×1 CBL is combined with the feature map of Path D through the cascade operation. The feature maps of combination and Path B are merged. After the merge operation, the channel shuffle is used as the output of FENB.

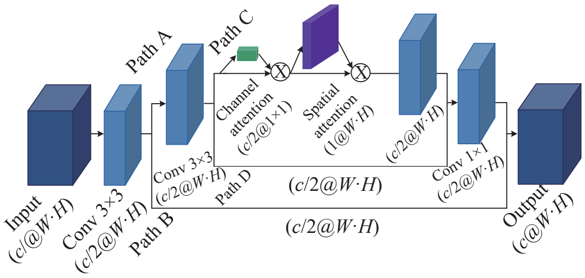
As shown in Figure 5a, the extracted features by convolution are very limited because of existing boundary effects. The channel shuffle operation (Figure 5b) readjusts feature map channel locations to enable information communication between the two Paths.



a. No channel shuffle b. Channel shuffle

Figure 5 Comparison with or without channel shuffle mechanism

Figure 6 shows the feature map width, height, and channel number of the FENB.



Note: H and W are the height and width of the feature map; @ is the split symbol.

Figure 6 Changes of width, height, and channel number of FENB feature map

FENB will be used to enhance the feature extraction ability of CSPDarknet53_tiny and improve the model accuracy. In FENB, when the feature maps are processed by convolution, the Memory Access Cost (MAC) is

$$\text{MAC}=HW(c_1+c_2)+c_1c_2 \tag{1}$$

$$B=HWc_1c_2 \tag{2}$$

where, H and W are the height and the width of the feature map; c_1 and c_2 are the numbers of input and output feature map channels. The mean value inequality can be inferred as

$$\text{MAC2}\sqrt{HWB}+\frac{B}{HW} \tag{3}$$

If and only if c_1=c_2, the MAC obtains the minimum value. Therefore, this study reduced the number of channels from c to c/2 through channel split instead of using convolution, so that the Memory Access Cost was minimum and the computing speed became faster.

2.2.3 Proposed algorithm

Figure 7 shows the model structure of the improved multi-class cherry tomatoes detection. The proposed model still applies CSPDarknet53_tiny as the model backbone network that is supplemented by Feature Enhancement Network (FEN). The FEN consists of three FENBs and two Down_samp modules. The Down_samp module consists of CBL with 1×1 convolution and Maxpool, which adjusts the number of channels and the size of feature maps.

The input images of 416×416×3 (height and width are 416, the number of channels is 3) are extracted by two CBL modules to extract the shallow information. The dimension is transformed to 104×104×64.

The shallow feature level of 52×52×128 is acquired from the feature map 104×104×64 by the first Resblock_body, which is extracted effective features by the first FENB obtaining the enhanced feature level of 52×52×128.

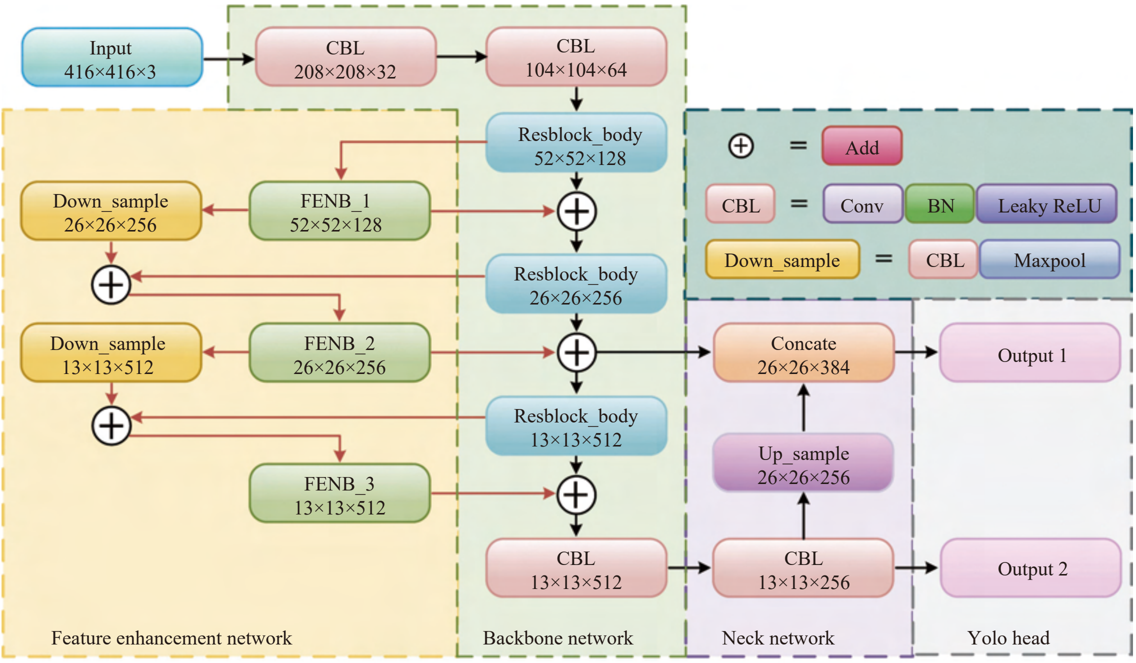
Then the results of the first Resblock_body and the first FENB are added as an input to the second Resblock_body. The first FENB output feature map is downsampled, dimensionally transformed to 26×26×256, and summed with the output 26×26×256 feature map of the second Resblock_body. Then, the 26×26×256 enhanced feature map is obtained through the second FENB and added to the output of the second Resblock_body, where the result feature map is served as the input to the neck network. By analogy, the third FENB and the second Down_samp are used to complete the feature enhancement of the third Resblock_body. Then the 13×13×512 feature map is obtained by CBL, which is used as the input feature map of the neck network. Considering the actual situation of picking a robot vision system to acquire images, the dimensions of 13×13×512 and 26×26×256 obtained by the Backbone Network, which complete feature fusion in the neck network to detect large and medium cherry tomatoes to simplify the model.

3 Experiments and discussion

3.1 Environment and parameters

The proposed model in this study was built and modified using the PyTorch framework. The training platform included a computer with Intel(R) Xeon(R) Silver 4210R CPU, a GPU of NVIDIA Quadro RTX 5000, and 16 GB of memory, running on a Windows 10 64-bit system. The software tools included CUDA10.1, CUDNN 7.6.4, Python 3.7, and Pycharm 2021.1.1.

In the experimental environment, the batch size was 8 and the number of threads was 4. This study trained 300 epochs to analyze



Note: FENB: Feature Enhancement Network Block.

Figure 7 Proposed model structure of multi-class cherry tomatoes detection algorithm.

the training process. Each epoch was trained using all the samples of the training set, where eight samples of one batch size were taken for one iteration. So, the proposed model was trained for 357 600 and 304 200 iterations on the training set of nighttime and daytime, respectively. The Adam optimizer was used to update and compute network. The momentum was 0.9. The decay weight was 0.0005. The learning rate was set to 0.001 for the first 50 epochs and 0.0001 for the last 250 epochs.

3.2 Evaluation of model performance

To accurately evaluate the performance of the detection model, Precision, Recall, Average Precision, mean Average Precision (mAP), and F1 score were used to examine the performance of the target detection algorithm.

According to the difference between the true classes and the predicted classes, all samples were divided into True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The definitions of Precision (P) and Recall (R) are as follows:

$$P=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{5}$$

where, P is the proportion of all positive predictions that are correct; TP represents the prediction is positive, and the ground truth is positive; FP represents the prediction is positive, and the round truth is negative; R is the proportion of all real positive observations that are correct; FN represents the prediction is negative, and the round truth is positive.

Average Precision (AP), mean Average Precision (mAP), and F1 score were used to evaluate the performance of model classification. They are defined as follows:

$$\mathrm{AP}_{\mathrm{class}}=\int_0^1 P_{\mathrm{class}}R_{\mathrm{class}}\mathrm{d}R_{\mathrm{class}} \tag{6}$$

$$\mathrm{mAP}=\frac{1}{n}\sum_1^n \mathrm{AP}_{\mathrm{class}} \tag{7}$$

$$\text{F1 score}=\frac{2\cdot P\cdot R}{P+R} \tag{8}$$

where, n=4 is the number of categories of the class; $\mathrm{AP}_{\mathrm{class}}$ is the AP value of different tomatoes; classes are NO, OF, OB, and OL; the mAP is the mean of AP of the four classes of cherry tomatoes; F1 score is the harmonic mean of P and R. This study mainly discussed AP and F1 score as the measurement standard for model performance.

3.3 Training evaluation and performance of the network

The training and validation loss curves of the proposed model are shown in Figure 8. The orange and blue curves are the loss curves of the training set and the validation set of the day, respectively. The red and green curves are the loss curves of the training set and the validation set of the night, respectively.

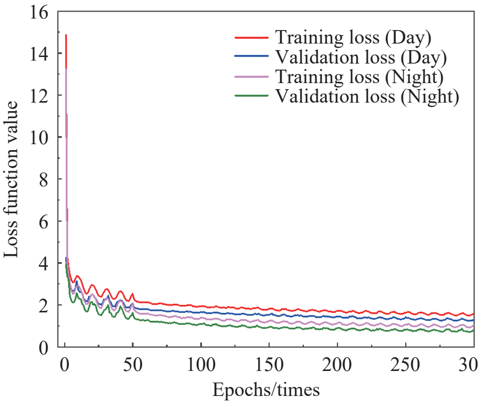


Figure 8 Loss curves of the improved models

As the number of epochs increases, all curve loss values generally decrease. In the first 50 epochs, values of all curves continued to fluctuate by the same amount after a rapid drop. The downward trend of the four curves was relatively stable after the 50th epoch. The four-curve loss values tended to stabilize at around the 280th epoch. The results show that the proposed model has no overfitting and good generalization ability. The parameters are appropriately selected.

The results of the *P-R* curve on the test set achieved by the improved multi-class detection model of cherry tomatoes are shown in Figure 9. The confidence was 0.5 in this study, where the results of *P*, *R*, AP_{class}, and F1 score are shown in Figure 10 under this level.

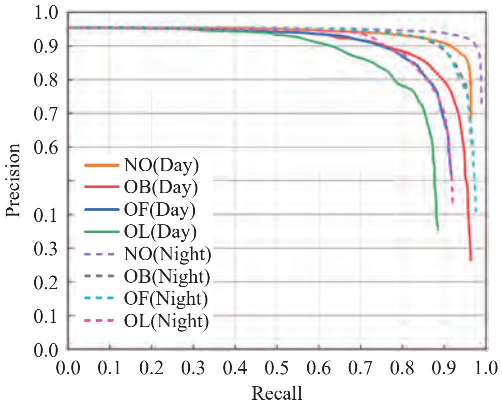
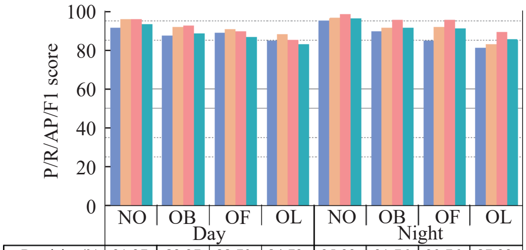


Figure 9 *P-R* curves of the proposed model on the testing data set



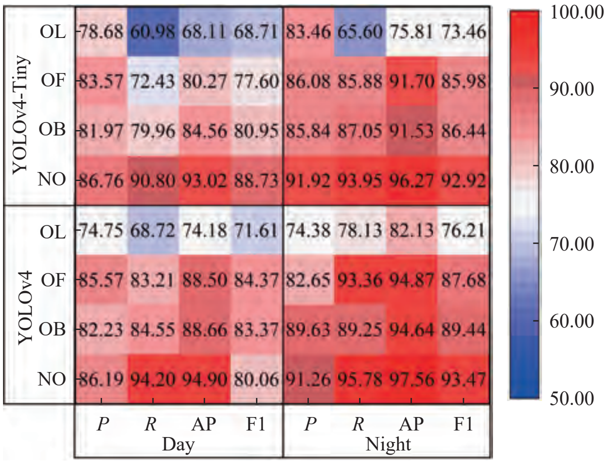
	Day NO	Day OB	Day OF	Day OL	Night NO	Night OB	Night OF	Night OL
■ Precision/%	91.37	83.37	88.70	84.73	95.93	91.76	90.76	87.99
■ Recall/%	95.02	89.56	84.61	81.11	96.40	91.41	91.67	82.87
■ AP/%	95.86	92.59	89.66	84.99	98.43	95.62	95.50	89.33
■ F1 score/%	93.16	88.45	86.61	82.88	96.16	91.58	91.21	85.35

Figure 10 Multi-class cherry tomatoes detection results with the proposed models

The proposed model has an obvious detection effect on the four classes of cherry tomatoes. As expected, the non-occluded cherry tomatoes can be easily detected as compared to the occluded fruits. Branch-occluded cherry tomatoes got the second-highest detection results on the day and night test sets. Since the branch was relatively thin, the degree of occlusion of target cherry tomatoes was relatively small to achieve higher detection metrics in the occluded classes. The AP and F1 score were second only to the fruits occluded by branch. Once the outer non-occluded cherry tomatoes are picked, the covered cherry tomatoes become unoccluded cherry tomatoes in subsequent detections. Leaf-occluded cherry tomatoes achieved the lowest detection results of AP and F1 scores, which were 84.99% and 82.88% on the day test set and 89.33%, and 85.35% on the night test set, respectively. The consequence of this class were significantly lower than others because the body of cherry tomatoes was small and the degree of occlusion of the leaves was different.

3.4 Comparison with classic target detection models

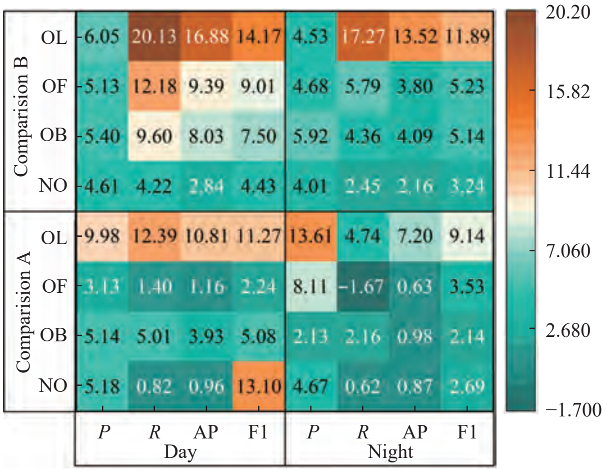
To verify the superiority of the proposed model, two one-stage classical models, YOLOv4 and YOLOv4-Tiny, were used to test the multi-class detection efficiency of cherry tomato images during the day and night. In this study, the relevant experimental parameters of the comparative models were strictly controlled, which was consistent with the proposed model. All models were run on the same training set, validation set, and test set. The results of the three models are shown in Figure 11 on the test set.



		Day *P*	Day *R*	Day AP	Day F1	Night *P*	Night *R*	Night AP	Night F1
YOLOv4-Tiny	OL	78.68	60.98	68.11	68.71	83.46	65.60	75.81	73.46
	OF	83.57	72.43	80.27	77.60	86.08	85.88	91.70	85.98
	OB	81.97	79.96	84.56	80.95	85.84	87.05	91.53	86.44
	NO	86.76	90.80	93.02	88.73	91.92	93.95	96.27	92.92
YOLOv4	OL	74.75	68.72	74.18	71.61	74.38	78.13	82.13	76.21
	OF	85.57	83.21	88.50	84.37	82.65	93.36	94.87	87.68
	OB	82.23	84.55	88.66	83.37	89.63	89.25	94.64	89.44
	NO	86.19	94.20	94.90	80.06	91.26	95.78	97.56	93.47

Figure 11 Performance indicators of different models on the same test set during the day and night

In Figure 12, comparison A shows the difference of related indicators between the proposed model and the YOLOv4 for the classification of the four cherry tomatoes at day and night, and comparison B shows the difference of related indicators between the proposed model and YOLOv4-tiny for the classification of the four cherry tomatoes at day and night.



		Day *P*	Day *R*	Day AP	Day F1	Night *P*	Night *R*	Night AP	Night F1
Comparision B	OL	6.05	20.13	16.88	14.17	4.53	17.27	13.52	11.89
	OF	5.13	12.18	9.39	9.01	4.68	5.79	3.80	5.23
	OB	5.40	9.60	8.03	7.50	5.92	4.36	4.09	5.14
	NO	4.61	4.22	2.84	4.43	4.01	2.45	2.16	3.24
Comparision A	OL	9.98	12.39	10.81	11.27	13.61	4.74	7.20	9.14
	OF	3.13	1.40	1.16	2.24	8.11	-1.67	0.63	3.53
	OB	5.14	5.01	3.93	5.08	2.13	2.16	0.98	2.14
	NO	5.18	0.82	0.96	13.10	4.67	0.62	0.87	2.69

Figure 12 Comparison of related indicators of different models

As expected, all models achieved the best detection performance on the non-occluded fruit class. For the detection of branch-occluded fruit, the advantages of AP and F1 scores of the proposed model are the most obvious, which were 3.93% and 5.08% higher than YOLOv4 and 8.03% and 7.50% higher than YOLOv4-Tiny on the day test set, respectively. The proposed model also achieved a similar improvement in the detection index of fruit occlusion as the branch occlusion. Leaf-occluded fruit obtained the worst detection results among the four categories based on the three models. However, the proposed model had the highest AP and F1 scores among the three models, which were 10.81% and 11.27% higher for the daytime dataset and 7.20% and 9.14% higher for the night dataset than YOLOv4, respectively. The AP and F1 scores of the improved algorithm were obviously improved by 16.88% and 14.17% for the daytime test set and 13.52% and 11.89% for the night test set, respectively. The proposed model reached the mAP of 90.78% and 94.72% on the day and night test set, which were 4.22% and 2.42% higher than YOLOv4 and 9.29% and 5.89% higher than YOLOv4-Tiny, respectively.

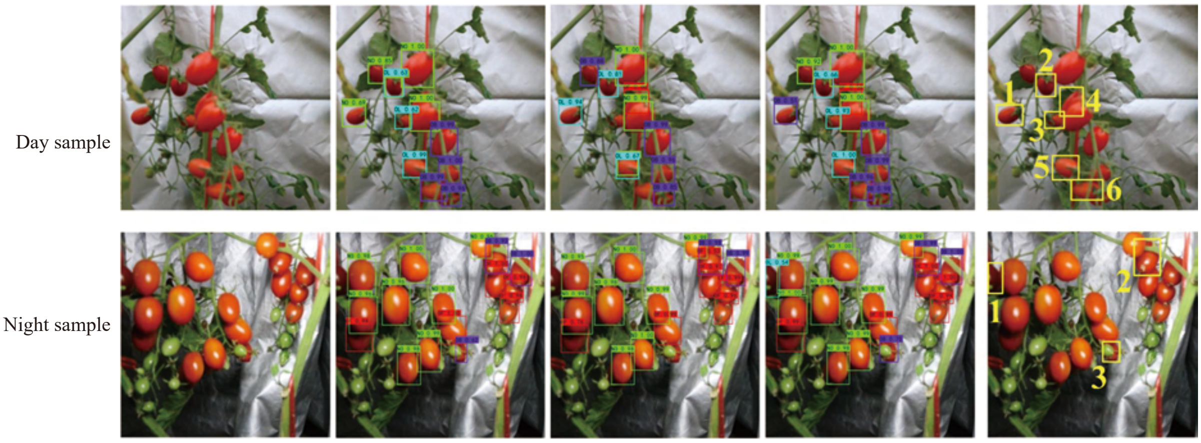
To sum up, the proposed model was the most accurate for detecting the different classes of cherry tomatoes during the day and night. The reason is that FENB integrated, extracted, and paid more attention to the characteristic of the detected cherry tomatoes on the one hand. On the other hand, the proposed model fused the enhanced feature maps with the original network features by means of FEN to strengthen the network object recognition ability. In addition, the mAP of the day was lower than at night for the three models, indicating that artificial LEDs provide good and reliable lighting conditions for the night-time identification of cherry tomatoes in the picking robot vision system. Table 1 lists the other test results of the three models.

Table 1 Results of the three models

Model		YOLOv4	YOLOv4-Tiny	Proposed Model
mAP/%	Day	86.56	81.49	90.78
	Night	92.30	88.83	94.72
Averagerun-time/ms		51.77	31.54	32.22
Model size/MB		245.53	23.10	39.34
Numberof parameters		6.44×10^7	0.61×10^7	1.03×10^7

The proposed model obviously outperformed YOLOv4 in terms of time-consuming, model size, and the number of parameters. Although the model size and the number of parameters of this method were 1.70 times and 1.69 times that of YOLOv4-Tiny, respectively, the average recognition time of a single image is only 1.02 ms longer than it. This effect had largely the benefit of the truncated gradient flow of CSPNet. Apart from this, the P, R, AP, mAP, and F1 scores were significantly better than YOLOv4 and YOLOv4-Tiny. Therefore, the proposed model achieved the expected detection performance, which can be adapted to the deployment of mobile terminals or embedded devices to meet the requirements of agricultural robot target detection systems.

The results from the corresponding models are shown in Figure 13. For the day sample, YOLOv4 and YOLOv4-Tiny missed detection of OL, OF, and OB with severe occlusion at 1, 4, and 6, which also repeated detection and misclassification at 2 and 5. The method proposed in this study acquired accurate detection at 2, 3, 4, 5, and 6, but incorrectly identified NO as OB at 1. For the night sample, neither YOLOv4 nor YOLOv4-Tiny detected OL at 1. YOLOv4-Tiny repeatedly detected OL into OL and OF at 2. YOLOv4 did not detect OB at 3. However, the proposed model accurately achieved detection at 1, 2, and 3. From these experiments, one can conclude that the proposed model has a significant effect on detecting multi-class cherry tomatoes during the day and night. The FENB efficaciously extracted cherry tomatoes features of different occlusion classes.



a. Original b. YOLOv4 c. YOLOv4-Tiny d. Proposed model e. Manual annotations

Note: The rectangles of green, blue, red, and purple colors are referring to the detected non-occluded fruit, leaf-occluded fruit, fruit-occluded fruit, and branch-occluded fruit, respectively. The yellow rectangles in Figure 13e are manual annotations indicating the difference between the proposed model and YOLOv4 and YOLOv4-Tiny.

Figure 13 Day and night examples comparing the detection effects of the improved model and the other two models

4 Conclusions

In this study, the CSPNet structure with the hybrid attention mechanism was developed to construct Feature Enhancement Network Block (FENB) for extracting efficient features. FENB was used to build Feature Enhancement Network to improve the detection accuracy of YOLOv4-Tiny. The experimental results showed that, under the confidence of 0.5, the average precision of non-occluded, branch-occluded, fruit-occluded, and leaf-occluded fruit over the day test images were 95.86%, 92.59%, 89.66%, and 84.99%, respectively, which were 98.43%, 95.62%, 95.50%, and 89.33% on the night test images, respectively. The mAP of four classes over the night test set was higher (94.72%) than that of the day (90.78%), which were both better than YOLOv4 and YOLOv4-Tiny. It cost 32.22 ms to process a 416×416 image on the GPU. The model size was 39.34 MB. Therefore, the proposed model provided a practical and feasible method for the multi-class detection of cherry tomatoes.

Acknowledgments

This work was financially supported by National Natural Science Foundation of China (Grant No. 52075149), Frontier Exploration Projects of Longmen Laboratory (Grant No. LMQYTSKT032), Scientific and Technological Project of Henan Province (Grant No. 212102110029), High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province (Grant No. JNZ201901), Colleges and Universities of Henan Province Youth Backbone Teacher Training Program (Grant No. 2017GGJS062), and Postgraduate Education Reform Project of Henan Province (Grant No. 2021SJGLX005Y; 2019SJGLX063Y).

[References]

[1] Rong J C, Wang P B, Yang Q, Huang F. A field-tested harvesting robot for oyster mushroom in greenhouse. Agronomy, 2021; 11(6): 1210.

[2] Zhang F, Chen Z J, Wang Y F, Bao R F, Chen X G, Fu S L, et al. Research on flexible end-effectors with humanoid grasp function for small spherical fruit picking. Agriculture, 2023; 13(1): 123.

[3] Rong J C, Wang P B, Wang T J, Ling H, Yuan T. Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots. Computers and Electronics in Agriculture, 2022; 202: 107430.

[4] Afroza A, Ambreen N, Baseerat A; Nigeena N, Ahmad S P, Azrah I S, Amreena S; Insha J, Majid R. Evaluation of Cherry Tomato (*Solanum lycopersicum* L. var. cerasiforme) Genotypes for Yield and Quality Traits. Journal of Community Mobilization and Sustainable Development, 2021; 16(1): 72–76.

[5] Yamamoto K, Guo W, Ninomiya S S. Node detection and internode length estimation of tomato seedlings based on image analysis and machine learning. Sensors, 2016; 16(7): 1044.

[6] Wang Z L. Underwood J. Walsh KB. Machine vision assessment of mango orchard flowering. Computers and Electronics in Agriculture, 2018; 151: 501–511.

[7] Wu J G, Zhang B H, Zhou J, Xiong Y J, Gu B X, Yang X L. Automatic Recognition of Ripening Tomatoes by Combining Multi-Feature Fusion with a Bi-Layer Classification Strategy for Harvesting Robots. Sensors, 2019; 19(3): 612–612.

[8] Xiong J T, Lin R, Liu Z, He Z L, Tang L Y, Yang Z G Zou X J. The recognition of litchi clusters and the calculation of picking point in a nocturnal natural environment. Biosystems Engineering, 2018; 166: 44–57.

[9] Bechar A, Vigneault C. Agricultural robots for field operations: Concepts and components. Biosystems Engineering, 2016; 149: 94–111.

[10] Silwal A, Davidson J R, Karkee M, Mo C K, Zhang Q, Lewis K. Design, integration, and field evaluation of a robotic apple harvester. Journal of Field Robotics, 2017; 34(6: 1140–1159.

[11] Silwal A, Karkee M, Zhang Q. A hierarchical approach to apple identification for robotic harvesting. Transactions of the ASABE, 2016; 59(5): 1079–1086.

[12] Guo Y M, Liu Y, Oerlemans A, Lao S Y, Lew M S. Deep learning for visual understanding: A review. Neurocomputing, 2016; 187: 27–48.

[13] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; 770–778.

[14] Liu F, Liu Y K, Lin S, Guo W Z, Xu F, Zhang B. Fast recognition method for tomatoes under complex environments based on improved YOLO. Transactions of the Chinese Society for Agriculture Machinery, 2020; 51(6): 229–237. (in Chinese)

[15] Vougioukas S G. Annual Review of Control, Robotics, and Autonomous Systems. Annual Reviews, 2019; 2: 365–392.

[16] Zhang F, Chen Z J, Bao R F, Zhang C C, Wang Z H. Recognition of dense cherry tomatoes based on improved YOLOv4-LITE lightweight neural network. Transactions of the Chinese Society of Agricultural Engineering, 2021; 37(16): 270–278. (in Chinese)

[17] Xu Z F, Jia R S, Liu Y B, Zhao C Y, Sun H M. Fast Method of Detecting Tomatoes in a Complex Scene for Picking Robots. IEEE Access, 2020; 8: 55289–55299.

[18] Zhang W J, Zhao X X, Ding R R, Zhang Z, Jiang H H, Liu P Z. A Detection and Recognition Method for Tomato on Faster R-CNN Algorithm. Journal of Shandong Agricultural University (Natural Science Edition), 2021; 52(4): 624–630. (in Chinese)

[19] Xu C, Xiong Z, Jiang X P, Deng M, Huang G C. Design and research of the cluster tomato picking robot. Modern Agricultural Equipment, 2021; 42(6): 15–23. (in Chinese)

[20] Zhang Q, Liu F P, Jiang X P, Xiong Z, Xu C. Motion planning method and experiments of tomato bunch harvesting manipulator. Transactions of the CSAE, 2021; 39(7): 149–156. (in Chinese)

[21] Gao F F, Fu L S, Zhang X, Majeed Y, Li R, Karkee M, et al. Multi-class fruit-on-plant detection for apple in SNAP system using Faster RCNN. Computers and Electronics in Agriculture, 2020; 176: 105634.

[22] Suo R, Gao F F, Zhou Z X, Fu L X, Song Z Z, Dhupia J, et al. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Computers and Electronics in Agriculture, 2021; 182: 106052.

[23] Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 39(6): 1137–1149.

[24] Zhao D A, Wu R D, Liu X Y, Zhao Y Y. Apple positioning based on YOLO deep convolutional neural network for picking robot in complex background. Transactions of the CSAE, 2019; 35(3): 164–173. (in Chinese)

[25] Wang C, Bochkovskiy A, Liao H M. Scaled-YOLOv4: Scaling cross stage partial network. In: 20221 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021; pp.13024–13033. doi: 10.1109/CVPR46437.2021.01283.

[26] Li H P, Li C Y, Li G B, Chen L X. A real-time table grape detection method based on improved YOLOv4-tiny network in complex background. Biosystems Engineering, 2021; 212: 347–359.

[27] Xu B, Wang N Y, Chen T Q, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint, 2015; arXiv: 1505.00853.

[28] Zheng Z H, Wang P, Liu W, Li J Z, Ye R G, Ren D W. Distance-IoU loss: Faster and better learning for bounding box regression. arXiv preprint, 2020; arXiv: 1911.08287.

[29] Wang C Y, Liao H M, Wu Y H, Chen P Y, Hsieh J W, Yeh I H. CSPNet: A new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020; pp.1571–1580.

[30] Woo S, Park J C, Lee J, Lweon I. CBAM: Convolutional Block Attention Module. In:Computer Vision - ECCV, 2018; 3–19.